Supporting Information for
Using pattern recognition of epigenetic signals for supervised enhancer prediction

Methods

Creation of Metaprofile:

We utilized the smoothed histone signal tracks provided for the S2 cell-line by the modENCODE consortium [1] to aggregate the corresponding histone signals around the STARR-seq peaks [2]. This aggregation was performed to remove noise before using the metaprofile *s(n)* for identifying active regulatory regions in the genome. The genome-wide profile for open chromatin (DNase-seq or DHS) for the S2 cell-line was calculated based on the experiments by the Stark lab [2]. To create the smoothened metaprofile, we aggregated the H3K27ac signal of active STARR-seq peaks with a noticeable “double peak” pattern within the H3K27ac signal in the S2 cell-line. All the STARR-seq peaks that overlap with DHS or H3K27ac peaks are assumed to be active regulatory regions in the genome.

To identify double peak regions, we initially identified the minimum in the H3K27ac signal track closest to the middle of the STARR-seq peaks. A minimum is accepted if it has the lowest signal within a 100 base pair region in the H3K27ac signal track. Then we proceed to identify the flanking maxima (both sides of the minimum) within a total of 2-kilo base pair region of the STARR-seq peak (1kb on each direction from the center of the STARR-seq peak). These maxima are accepted only if they have the highest signal within a 100 base pair region in the H3K27ac signal track. Approximately 70% of the active STARR-seq peaks contained an identifiable double peak within the H3K27ac signal.

After identifying the double peaks surrounding STARR-seq peaks, we aggregated the signal after aligning the maxima flanking the regulatory region. The signal track is interpolated with a cubic spline fit so that the signal track contains equal number of points for each double peak region. All interpolation and smoothing steps were performed using the scipy module in python. The aggregated signal tracks are averaged to create the metaprofile for the double peak regions. While the signal tracks are aggregated based on identifying the double peak regions in the H3K27ac signal track, the same set of operations can be performed with any epigenetic mark expected to have the double peak pattern flanking regulatory regions.

In addition, while creating the metaprofile for H3K27ac signal close to active STARR-seq peaks, we also performed the same set of transformations on other dependent epigenomic datasets (other histone marks and/or DHS signal). In this study (Figures 1 and S2), the dependent profiles for all other epigenetic datasets are calculated by averaging the corresponding signal based on identifying double peak regions within H3K27ac signal. If the signal tracks of the other epigenetic marks also tend to contain a double peak pattern in the same regions, the metaprofiles for the corresponding epigenetic marks will also contain a double peak pattern as observed in Figure S2A. However, as DHS and repressive histone marks do not contain a double peak pattern (Figure S2), these regions do not have the same epigenetic template associated with enhancers.

Matched Filter Algorithm:

The epigenetic signal at enhancers and promoters can be approximated as the linear superposition of background noise and the metaprofile *s(n)* learned in Figure 1 (Figure S2) for the corresponding experimental dataset. The matched filter *h(n)* is used to scan the epigenetic signal to identify the occurrence of the metaprofile pattern within different regions of the genome. Before calculating the matched filter score, interpolation of signal is used to ensure that the scanned region contains the same number of points as the metaprofile. The matched filter process is equivalent to the computation of the cross correlation between the signal *y(n)* and the reverse of the transformed metaprofile template *s*(N-n)* (where *N* is the total number of points in the template). In other words:

$$r(n) = \sum_{i=1}^{N} y(i) * h(i)$$

where *h(i)* is the matched filter and can be written as:

$$h(i) = s^{*}(N - i)$$

As shown in Figure S1, there is a large amount of variability in the span (distance between the two peaks in the histone signal) of the regulatory region in the epigenetic signal. As a result, we scan the genome with the matched filter scanning different spans of the genome (distance between the two peaks allowed to vary between 300 and 1100 base pairs) and take the highest score as the matched filter score for that region. The matched filter is the filter that recognizes any given template in the presence of noise in a signal with the highest signal-to-noise ratio [3]. In the presence of white noise alone, the matched filter score is low and follows a Gaussian distribution (negatives). The presence of the metaprofile within the signal leads to higher matched filter scores for positives.

Statistical Learning Models

The matched filter scores for negatives for different histone marks are unimodal that can be fit using separate Gaussian distributions. The Z-scores of matched filter scores with respect to the negatives (random regions of genome) are used as input features for training different statistical learning models. The Z-score of the matched filter score for a region (*z(i)*) is:

$$z(i) = \frac{r(i) - \mu}{\sigma}$$

where *r(i)* is the matched filter score for region *i* while μ *and* σ are the mean and standard deviation of the Gaussian fit to the matched filter scores for random regions in genome. In the main text, we discuss our results of the Support Vector Machine (SVM) model, which is one of the most versatile and successful binary classifiers [4]. We utilized a linear kernel to distinguish between the positives and negatives. The linear SVM identifies a decision boundary that maximally discriminates the epigenetic features of regulatory regions from random regions of the genome in the SVM feature vector space.

In Figure S5, we also present results for Ridge Regression [5], Random Forest [6], and Gaussian Naïve Bayes [7] models and the accuracy of different models are comparable.

Ridge regression is a linear regression technique that prevents over fitting by penalizing large weights for each feature. Random Forest is an ensemble learning method that operates by constructing a large number of decision trees and outputting the mean prediction of different decision trees. We used thousand trees for creating our enhancer and promoter prediction models. The naïve Bayes classifier is a family of simple probabilistic classifiers that assumes that all the features are independent of one another. We used scikit-learn [8] with default parameters for training and assessing the performance of all the statistical models. In general, the SVM and random forest models performed the best over all the tests and were the most flexible models.

Assessing the Models:

In order to assess the accuracy of matched filter for predicting enhancers and promoters, we used 10-fold cross validation. During 10-fold cross validation, the positives and negatives are randomly divided in to 10 groups each. Nine of the 10 groups are randomly combined to train the model and the predictions are tested on the 10th group. To evaluate the performance of trained classifiers, we performed 10-fold cross-validation on the training data and quantified our results with area under receiver-operating characteristic (ROC), and area under precision-recall (PR) curves.

In the ROC curve [9], the true positive (TP) rate is plotted against the false positive (FP) rate at different thresholds in the statistical model. The TP rate is defined as the fraction of positives identified correctly by the model (i.e., ratio of number of true positives identified by the model to the total number of positives). The FP rate is defined as the fraction of negatives identified correctly by the model (i.e., ratio of number of negatives misclassified by the model to the total number of negatives). While comparing the performance of two different classifiers in the ROC curve, the classifier with higher TP rate at the same FP rate is considered to be a better classifier. The area under the ROC is a single measure for the accuracy of a model as models with higher area under ROC are generally considered to be better models.

In the PR curve, the precision is plotted against recall at different thresholds in the statistical model. The recall is the same as the TP rate of the model (i.e., ratio of number of true positives identified by the model to the total number of real positives). The precision is the fraction of positives in the model that are correct (i.e., ratio of number of true positives identified by the model to the total number of positives according to the model). In skewed datasets with large number of negatives in comparison to positives, the FP rate can be low even when the number of false positives misclassified by the model is comparable to the number of true positives. For such skewed datasets, te area under ROC for two different models may be very similar even though they actually differ in performance with respect to their precision. Hence, the area under the PR curve is a better reflection of the performance difference between two models with similar area under ROC in skewed datasets.

In Figure 2, the positives are defined as the active peaks (intersecting with DHS or H3K27ac peaks) from a single STARR-seq experiment (singe core promoter) or the union of active peaks from multiple STARR-seq experiments (multiple core promoters). The negatives are randomly chosen regions in the genome with H3K27ac signal that had the same width distribution as the distribution of distance between double peaks near STARR-seq peaks (shown in Figure S1). We typically chose between 5 to 10x number of negatives as compared to number of positives in Figures 2, 3, and 4 as the number of enhancers and promoters in the genome (positives) are far lesser than the number of negatives and area under PR curve is dependent on the ratio of negatives to positives during 10-fold cross validation. The matched filter score for each region is chosen as the best matched filter score with a 1500 bp region centered on each positive and negative. The matched filters are scanned with distances between 300-1100 bp before choosing the best score. While comparing the performance of the matched filter to the peak-based models of the different epigenetic marks (Figure S4), we assumed that histone (DHS) peaks that overlapped with at least 50% (10%) of the STARR-seq peak is used to rank that prediction. We used a smaller threshold for DHS peaks as they are much smaller than histone peaks. We achieved similar results with thresholds of 25% for both histone and DHS peaks. The p-value of the intersecting peak is used to rank the peak-based predictions. The modENCODE histone peaks [1] and DHS peaks [2] were compared to the matched filter scores in Figure S4.

During STARR-seq, each peak is functioning as an enhancer within the plasmid environment in S2 cell-line. However, to delineate the native role of the region, we classify them as promoters and enhancers based on their distance to the transcription start sites in the genome. In Figure 3, the active promoters are defined as active STARR-seq peaks (multiple core promoter) within 1 kb of TSS (Ensembl release 78) while enhancers were active STARR-seq peaks more than 1kb from any TSS in *Drosophila melanogaster*. While calculating the matched filter for positives and negatives, we considered the best scoring matched filter score after padding each region to 1.5kb width.

In Figure 4, the promoters are defined as FIREWACh peaks within 2 kb of TSS (GENCODE release vM4) while enhancers were FIREWACh peaks more than 2kb from any TSS. The larger distance (2 kb) for defining promoters was used because of the larger size of the mouse genome. The FIREWACh assay is performed in a transduction assay and was based on ChIP-seq peaks of a few key TFs. Hence, we did not split the FIREWACh peaks in to active and poised enhancers and promoters. The ENCODE histone and DHS datasets for mESC were used to predict enhancers and promoters in Figure 4.

H1-hESC whole genome prediction

To predict enhancers and promoters on the whole genome, we utilized the 6 parameter machine learning model shown in Figure 2. The histone and DHS signals from ENCODE consortium [10] were used to predict enhancers and promoters in H1-hESC. The histone signals were converted to log fold enrichment (with respect to control signal) before we scanned it with the matched filter. There were 43463 active regulatory regions predicted in the human genome (< 2% of genome). All regions within 2kb of TSS were annotated as promoters while active regulatory regions that were more than 2kb from TSS were annotated as enhancers. The distribution of the expression of closest gene (GENCODE v19 TSS) from ENCODE RNA-seq dataset [10] for H1-hESC was compared to the expression of all genes from H1-hESC. The Wilcoxon test was used to measure the significance of changes in gene expression.

H1-hESC TF binding

To measure the differences in TF binding and co-binding patterns at promoters and enhancers, we overlapped the ChIP-seq peaks from ENCODE with our predicted enhancers and promoters using intersectBed. The two regions were considered to be overlapping if at least 25% of the ChIP-seq peak was overlapping with the predicted enhancer or promoter.

Table S1 – Performance of matched filter models with single epigenetic feature for all STARR-seq peaks (multiple core promoters)

Feature	AUROC	AUPR
H3K27ac	0.95	0.90
H3K4me1	0.70	0.59
H3K4me2	0.91	0.79
H3K4me3	0.84	0.76
H3K9ac	0.92	0.85
H4K12ac	0.92	0.86
H3	0.80	0.70
H1	0.88	0.81
H2BK5ac	0.94	0.90
H4K8ac	0.88	0.79
H4K5ac	0.87	0.79
H4K16ac	0.89	0.72
H3K18ac	0.90	0.84
H3K9me1	0.71	0.61
H3K79me2	0.79	0.58
H4K27me2	0.81	0.68
H2Av	0.66	0.57
H3K27me3	0.83	0.64
H3K23ac	0.66	0.46
H3K79me3	0.70	0.51
H3K27me1	0.64	0.43
H4	0.67	0.49
H3K36me1	0.54	0.41
H3K9me3	0.59	0.42
H3K9me2	0.60	0.41
H3K36me3	0.57	0.38
H4K20me1	0.47	0.31
H3K79me1	0.47	0.30

Table S2 – Performance of matched filter models with single epigenetic feature for promoters and enhancers (multiple core promoters). Numbers within (outside) parenthesis are accuracy of models for predicting promoters (enhancers).

Feature	AUROC	AUPR
H3K27ac	0.91 (0.96)	0.60 (0.73)
H3K4me1	0.88 (0.60)	0.42 (0.16)
H3K4me2	0.84 (0.92)	0.21 (0.48)
H3K4me3	0.62 (0.92)	0.09 (0.65)
H3K9ac	0.85 (0.94)	0.24 (0.70)
H4K12ac	0.90 (0.93)	0.33 (0.58)
H3	0.78 (0.83)	0.26 (0.48)
H1	0.83 (0.92)	0.36 (0.61)
H2BK5ac	0.91 (0.96)	0.59 (0.70)
H4K8ac	0.90 (0.86)	0.55 (0.37)
H4K5ac	0.89 (0.86)	0.52 (0.41)
H4K16ac	0.90 (0.90)	0.52 (0.40)
H3K18ac	0.90 (0.88)	0.60 (0.47)
H3K9me1	0.53 (0.81)	0.09 (0.44)
H3K79me2	0.70 (0.83)	0.10 (0.27)
H4K27me2	0.68 (0.85)	0.19 (0.44)
H2Av	0.63 (0.78)	0.15 (0.36)
H3K27me3	0.81 (0.86)	0.20 (0.36)
H3K23ac	0.55 (0.71)	0.07 (0.20)
H3K79me3	0.61 (0.74)	0.08 (0.23)
H3K27me1	0.72 (0.57)	0.12 (0.12)
H4	0.69 (0.68)	0.13 (0.21)
H3K36me1	0.75 (0.58)	0.19 (0.18)
H3K9me3	0.59 (0.64)	0.11 (0.15)
H3K9me2	0.62 (0.63)	0.09 (0.15)
H3K36me3	0.60 (0.62)	0.09 (0.14)
H4K20me1	0.55 (0.50)	0.07 (0.10)
H3K79me1	0.54 (0.58)	0.06 (0.12)

Figure Captions:

Figure S1: Variability in double peak pattern. A) The frequency of distance between the two maxima in a double peak flanking active STARR-seq peaks is plotted. B) The symmetricity of the double peak pattern is plotted. The ratio of the distance between the two peaks to the ratio between one of the maxima and the minima is plotted. While there is large amount of variability in the distance between the two peaks (mostly between 300-1100 bp), the trough in the double peak tends to occur in the center of the two peaks.

Figure S2: Metaprofile for different epigenetic marks. The metaprofile around active STARR-seq peaks is plotted for different epigenetic marks. Histone marks that are enriched near STARR-seq peaks display the characteristic double peak pattern shown in A) due to the depletion of histone proteins at active regulatory regions. In addition, DHS displays a single peak at the center of these regulatory regions as shown in A). B) On the other hand, no such double peak pattern is observed on depleted histone marks at STARR-seq peaks.

Figure S3: Histogram of matched filter scores. The probability density of matched filter scores for different epigenetic marks for STARR-seq peaks (positives) and random regions of the genome (negatives) with H3K27ac signal. In most cases, the matched filter scores for positives and negatives are Gaussian curves. The amount of overlap between these two curves determines the accuracy of the matched filter for predicting STARR-seq peaks using thematched filters for the corresponding epigenetic feature.

Figure S4: Accuracy of matched filter and peak-based models. The performance of the matched filters of different epigenetic marks and the peak-based models for predicting all STARR-seq peaks is compared here using 10-fold cross validation. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (multiple core promoters) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for the matched filter model. B) The individual ROC and PR curves for each matched filter and the peak-based model are shown.

Figure S5: Comparison of different statistical models. The performance of the different statistical models to integrate the information from six epigenetic features is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (single core promoter) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting STARR-seq peaks identified after combining multiple core promoters. B) The individual ROC and PR curves for each statistical model. C) The contribution of the matched filter score for each epigenetic feature to the different integrated models.

Figure S6: Comparison of different statistical models for 30-feature model. The performance of the different statistical models to integrate the information from 30 epigenetic features is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (single core promoter) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting STARR-seq peaks identified after combining multiple core promoters. B) The individual ROC and PR curves for each statistical model. C) The contribution of the matched filter score for each epigenetic feature to the different integrated models.

Figure S7: Histogram of matched filter scores for chosen features in promoters and enhancers. A) The histogram of matched filter scores for small set of epigenetic features on

promoters is compared to random regions of the genome. B) The histogram of matched filter scores for small set of epigenetic features on enhancers is compared to random regions of the genome.

Figure S8: Comparison of different statistical models for predicting enhancers and promoters. The performance of the different statistical models to integrate the information from six epigenetic features for promoter and enhancer prediction is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the promoters with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting enhancers. The promoters and enhancers from multiple STARR-seq experiments with different core promoters are merged in this analysis. B) The individual ROC and PR curves for each statistical model is shown. The contribution of the matched filter score for each epigenetic feature to the different integrated models for promoter prediction (C) and enhancer prediction (D) are shown.

Figure S9: Comparison of different statistical models for predicting enhancers and promoters. The performance of the different statistical models to integrate the information from thirty epigenetic features for promoter and enhancer prediction is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the promoters with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting enhancers. The promoters and enhancers from multiple STARR-seq experiments with different core promoters are merged in this analysis. B) The individual ROC and PR curves for each statistical model is shown. The contribution of the matched filter score for each epigenetic feature to the different integrated models for promoter prediction (C) and enhancer prediction (D) are shown.

Figure S10: Accuracy of enhancer-trained matched filter and statistical models for promoter prediction. The performance of the enhancer-trained matched filters of different epigenetic marks and statistical models for predicting active promoters is compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

Figure S11: Accuracy of promoter-trained matched filter and statistical models for enhancer prediction. The performance of the promoter-trained matched filters of different epigenetic marks and statistical models for predicting active enhancers is compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

Figure S12: Transferability of models across cell-lines. The performance of the BG3-trained matched filters of different epigenetic marks and statistical models for predicting active promoters and enhancers are compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

Figure S13: Location of H1-hESC predictions. A) The probability density of the distance of the predicted promoter and enhancer from the closest TSS is shown. B) The location of the enhancers and promoters on genomic elements are shown. Promoters are defined as TSS +/- 2kb. All TSS, UTR, exons, introns, and intergenic elements are calculated based on GENCODE 19 definitions [11]. A regulatory region is considered to overlap with the elements if more than 50% of the matched filter region overlaps with the corresponding element in B.

Figure S14: Gene expression of closest gene. The distribution of gene expression of gene closest to the enhancer/promoters are plotted and compared to the gene expression of all genes in H1-hESC. A Wilcoxon test shows that P-value for differences in gene expression of genes close to enhancers and promoters are significantly higher than expression of all genes in H1-hESC (< 10^{-100} each).

Figure S15: Overlap of TF binding site with predicted promoters/enhancers. The fraction of promoters and enhancers that overlap with different TF ChIP-seq peaks in H1-hESC are plotted. The color of the bar is plotted based on the fraction of ChIP-seq peaks for corresponding TF that overlap with the promoter/enhancer. The difference in patterns of TF binding was used to create models that distinguish enhancers from promoters (Figure 5B).

Figure S16: Patterns of co-TF binding on enhancers and promoters. The patterns of TF co-occurrence on a single matched filter prediction around promoters and enhancers are plotted. The differences between co-TF binding at enhancers and promoters can be used to gain some mechanistic insight into TF cooperativity.

References:

1. mod, E.C., et al., *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-97.
2. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
3. Kumar, V.B.V.K., A. Mahalanobis, and R.D. Juday, *Correlation Pattern Recognition*. 2005.
4. Blanchard, G., O. Bousquet, and P. Massaer, *Statistical performance of support vector machines.* Ann. Statist., 2008. **36**: p. 489-531.
5. Hoerl, A.E. and R.W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55--67.
6. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5--32.
7. Stuart, R. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2nd ed. 2003.
8. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825--2830.
9. Davis, J. and M. Goadrich, *The Relationship Between Precision-Recall and ROC Curves.* Proceedings of the 23rd international conference on Machine Learning, 2006: p. 233-240.
10. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
11. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.